Greenwich Academic Literature Archive (GALA)
– the University of Greenwich open access repository
http://gala.gre.ac.uk

___

___

___

Contact: gala@gre.ac.uk

SCIENTIFIC REP⊘RTS

The Scaling of Human Contacts and Epidemic Processes in Metapopulation Networks

Michele Tizzoni[1], Kaiyuan Sun[2], Diego Benusiglio[1,3], Márton Karsai[4] & Nicola Perra[2]

We study the dynamics of reaction-diffusion processes on heterogeneous metapopulation networks where interaction rates scale with subpopulation sizes. We first present new empirical evidence, based on the analysis of the interactions of 13 million users on Twitter, that supports the scaling of human interactions with population size with an exponent γ ranging between 1.11 and 1.21, as observed in recent studies based on mobile phone data. We then integrate such observations into a reaction- diffusion metapopulation framework. *We* provide an explicit analytical expression for the global invasion threshold which sets a critical value of the diffusion rate below which a contagion process is not able to spread to a macroscopic fraction of the system. In particular, we consider the Susceptible-Infectious-Recovered epidemic model. Interestingly, the scaling of human contacts is found to facilitate the spreading dynamics. This behavior is enhanced by increasing heterogeneities in the mobility flows coupling the subpopulations. Our results show that the scaling properties of human interactions can significantly affect dynamical processes mediated by human contacts such as the spread of diseases, ideas and behaviors.

The network of social interactions between individuals in a community represents the main substrate for a number of spreading phenomena such as the diffusion of infectious diseases, ideas and behaviors[1–3]. In the past fifteen years network science has developed a wide range of mathematical tools to study and model such dynamical processes[3–5]. In particular, building upon a long research tradition in ecology[6], the theoretical framework of reaction-diffusion (RD) processes on metapopulation networks has been proved to be extremely valuable for describing contagion phenomena in spatially structured systems[7]. In this framework, individuals are represented by particles that reside in nodes of a network and migrate along the connections between them. Each node describes a subpopulation, i.e. a city or a town, while each link represents a travel route. Inside each node, particles react according to the rules of the process under study. Such modeling approach has been widely used to describe the dynamics of a number of real world complex systems[8–10]. Its most successful application, though, has been the modeling of the spread of infectious diseases in structured populations[11–24]. A common assumption in RD metapopulation models is that particles interact in each node with the same contact rate, constant and equal for any given size of the subpopulation. In mathematical epidemiology, such assumption corresponds to the frequency-dependent transmission rate[25]. However, in the past years, considerable efforts have been devoted to quantitatively measure human mixing patterns in a variety of settings, from small spatial and temporal scales[26] to country wide studies[27]. This has been possible thanks to the availability of new emerging technologies[28], such as RFID sensors[26], mobile phones[29] and social media[30]. A recent study based on the analysis of large mobile phone datasets[31] has shown evidence that the *per capita* social connectivity scales with the subpopulation size. In particular, the authors of[31] found that the cumulative

[1]Computational Epidemiology Laboratory, ISI Foundation, via Alassio 11/C, 10126, Torino, Italy. [2]Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston MA 02115 USA. [3]Dipartimento di Fisica, Universitá degli Studi di Torino, via Giuria 1, 10126, Torino, Italy. [4]Laboratoire de l'Informatique du Parallélisme, INRIA-UMR 5668, IXXI, ENS de Lyon, 69364 Lyon, France. Correspondence and requests for materials should be addressed to M.T. (email: michele.tizzoni@isi.it)

number of social contacts of individuals in a city scales as N^γ where $\gamma > 1$ and N is the city's population. This finding is consistent with a number of scaling properties observed in cities such as wages, crime rates, infrastructure per capita[32,33] and with theoretical models of urban development[34,35].

In this work, we first present new empirical evidence, based on the analysis of human interactions on Twitter that supports the contacts scaling hypothesis. Then, we integrate such observation into a RD metapopulation framework characterized by realistic heterogeneities in the distribution of the number of connections per node and in traffic flows. In particular, we study a Susceptible-Infectious-Recovered (SIR) epidemic dynamics inside each subpopulation[36]. We provide an explicit analytical expression for the global invasion threshold that sets a critical value of the diffusion/mobility rate below which a contagion process is not able to spread to a macroscopic fraction of the system[17]. We show that the scaling of interaction rates with subpopulation size significantly alters the contagion dynamics leading to a lower critical value of the mobility rate. Interestingly, such variations are enhanced by increasing heterogeneities in mobility patterns coupling the subpopulations.

Results

The scaling of human contacts on Twitter.
We analyze the interactions between users of the micro-blogging platform Twitter in several countries. We considered two different geographical aggregations (see Material and Methods for more details). The first maps about 13 millions Twitter users into 2,371 census areas centered around major transportation hubs[12] in 205 countries. Such aggregation level has been used to model pandemic spread at the global scale[13,24]. The second maps about 4.6 million Twitter users into 1,344 metropolitan areas, across the USA and 31 European countries.

To extract the relation between contacts and population size, we follow the methods used by Schläpfer *et al.*[31]. In both aggregation levels we build the communication network through Twitter mention interactions (see Material and Methods). In our analysis, a link is placed between users A and B within a given census area if and only if A mentioned B and vice versa at least once. We calculate the cumulative degree $C = \sum_{i=1}^{S} c_i$, where c_i is the degree of user i and S is the number of users within a census area, and rescale it by the users' coverage $\frac{S}{N}$ to obtain $C_r = C\frac{N}{S}$, where N is the total population of a census area obtained from official sources[37–39]. The rescaling procedure corresponds to an extrapolation of the observed average nodal degree, $\langle c \rangle = \frac{C_r}{N}$, to the entire population of the census area[31], and effectively reduces fluctuations due to variations in coverage from city to city.

To test the scaling hypothesis, we fit the rescaled cumulative degree C_r to a power-law function of the population of the census area, in the form $\ln C_r = \gamma \ln N + \epsilon$, and compare the result against a null model, represented by a linear function of the population, $\ln C_r = \ln N + \hat{\epsilon}$, where ϵ and $\hat{\epsilon}$ are constant. In all cases the power-law function is found to be a better fit to the data than the linear regression, based on the adjusted R^2, and the difference between the exponent $\gamma > 1$ and the simple linear regression is statistically significant ($p < 1.001$, details in the Supplementary Information file). More specifically, we find, consistently with Schläpfer[31], that the rescaled cumulative degree C_r is characterized by a power-law relation with the population of the census areas, $C_r \propto N^\gamma$ with exponent $\gamma = 1.11 \pm 0.01$ considering basins and $\gamma = 1.20 \pm 0.02$ considering metropolitan areas (see Fig. 1). We also restrict our analysis of the Twitter dataset to the two aggregation levels in the USA and Europe. We find that the scaling behavior still holds, with the exponent γ in the same range, i.e. $\gamma = 1.15 \pm 0.01$ in the USA and $\gamma = 1.21 \pm 0.04$ in Europe considering census areas, and $\gamma = 1.16 \pm 0.02$ in the USA and $\gamma = 1.18 \pm 0.02$ in Europe considering metropolitan areas.

Similar results are obtained considering also the connections of a user in the whole network. In particular, when the total number of Twitter interactions C is calculated by assuming c_i to be the degree of user i in the entire network, c_i is no longer confined within the basin/metropolitan area boundary and the interactions between user i and users from other basin/metropolitan areas or users that are not geo-mappable are also taken into account (see Fig. S2 in the Supplementary Information file). In this case, we find $\gamma = 1.11 \pm 0.01$ in the US, $\gamma = 1.06 \pm 0.01$ in Europe and $\gamma = 1.11 \pm 0.01$ when considering all the basins in the world. For the case of metropolitan areas, we find $\gamma = 1.07 \pm 0.01$ in the US, $\gamma = 1.09 \pm 0.03$ in Europe and $\gamma = 1.08 \pm 0.02$ when combining all the metropolitan areas of the US and Europe together. In Table 1 we report a complete summary of the values of γ computed at all scales and aggregation levels.

Global invasion threshold and numerical simulations.
To study the effect of the scaling of contact rates in RD processes, we consider a metapopulation network of V nodes, and N individuals. Each node i has degree k_i, and population size $N_i(t)$. The degree describes the number of subpopulations connected to it. We adopt a degree-block approximation, assuming all the subpopulations of degree k to be statistically equivalent[15–17]. We denote the degree distribution of the network as $P(k)$. To describe the diffusion of individuals, we assume that the rate at which individuals leave a subpopulation is independent of its degree and equal to p. However, to reproduce the properties of real transportation networks[40], we consider heterogeneous distributions of degree and traffic flow. In particular, the diffusion rate of individuals between two nodes of degree k and k' is $d_{kk'} = p\frac{\omega_0(k'k)^\theta}{T_k}$, where T_k provides the necessary normalization,
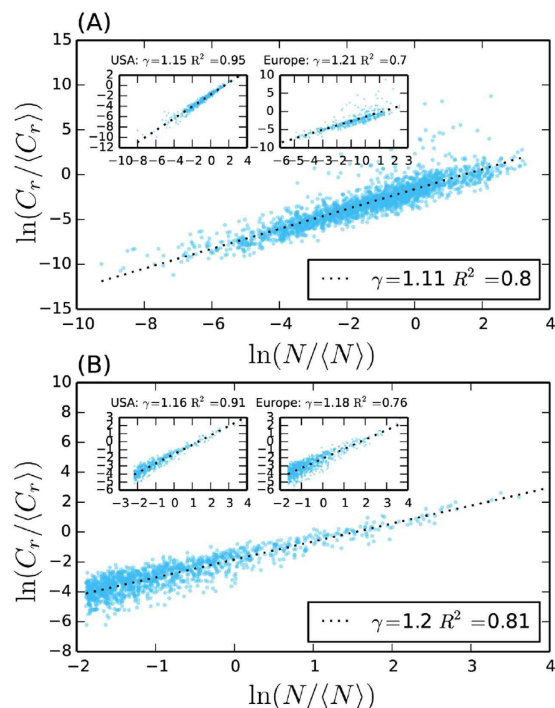
Figure 1. (**A**) Rescaled cumulative degree C_r against population N, measured between 13129406 Twitter users distributed across 2371 basins in 205 countries. (**B**) Rescaled cumulative degree against population, measured between 4606444 Twitter users in 1344 metropolitan areas in 31 countries. We normalized the values of C_r and N by their average to compare the results across different countries. Insets show the dependency of C_r on N restricted to the Twitter users in the US and Europe.

Geographical aggregation	Scaling exponent γ
Basins (internal connections only)	1.11 ± 0.01
Basins (all connections)	1.11 ± 0.01
Metro areas (internal connections only)	1.20 ± 0.02
Metro areas (all connections)	1.08 ± 0.02
Basins in US (internal connections only)	1.15 ± 0.01
Basins in US (all connections)	1.16 ± 0.02
Metro areas in US (internal connections only)	1.16 ± 0.02
Metro areas in US (all connections)	1.09 ± 0.03
Basins in Europe (internal connections only)	1.21 ± 0.04
Basins in Europe (all connections)	1.06 ± 0.01
Metro areas in Europe (internal connections only)	1.18 ± 0.02
Metro areas in Europe (all connections)	1.08 ± 0.02

Table 1. Summary of the scaling exponents γ measured on the Twitter dataset. Error intervals correspond to the standard error of the slope in the regression fit. When referring to Europe, the following 31 countries are taken into consideration: Belgium, France, Bulgaria, Bosnia Herzegovina, Croatia, Germany, Hungary, Finland, Denmark, Netherlands, Portugal, Latvia, Lithuania, Luxembourg, Romania, Poland, Greece, Estonia, Italy, Albania, Czech Republic, Cyprus, Austria, Ireland, Spain, Macedonia, Slovakia, Malta, Slovenia, United Kingdom, Sweden.

and ω_0 is a system dependent constant that rescales the diffusion rates between nodes. Without lack of generality we set $\omega_0 = 1$. It is possible to show that, under such conditions, the population of a node of degree k, $N_k(t)$, at equilibrium is given by $N_k = \bar{N}\frac{k^{1+\theta}}{\langle k^{1+\theta}\rangle}$, where $\bar{N} = \sum_k P(k)N_k$[17]. As a consequence, the exponent θ, which modulates the heterogeneity of the mobility flows, also regulates the heterogeneity

of the subpopulations size distribution. We model the reactions, taking place in each node, as a stochastic SIR epidemic process where individuals are partitioned according to their health status: susceptibles (S), infectious (I) and recovered (R). The SIR dynamics are defined by two transitions: the infection process $S + I \rightarrow 2I$, regulated by the transmissibility λ, and the recovery process $I \rightarrow R$, tuned by the recovery rate μ[2].

Here, we investigate the case in which the infection dynamics is dependent on the local population size. More precisely, inside each node, we consider an homogeneous mixing approximation where the average contact rate scales with the population size as $\bar{c} \sim N_i^\eta$. The values of the exponent γ measured in real social networks correspond to $\eta = \gamma - 1$ ranging between 0.11 and 0.2. The value of is η measured by Schläpfer *et al.*[31] is $\eta = 0.12$, 95% CI: $[0.11-0.15]$. Without lack of generality we focus on the case $\eta > 0$ in our analytical treatment, then we consider the range $\eta \in [0-0.48]$ for the numerical solutions of the system's equations and $\eta \in [0-0.12]$ for Monte Carlo simulations. The case $\eta = 0$ corresponds to the classic SIR model with frequency-dependent transmission rate while $\eta = 1$ corresponds to the density-dependent case. The immediate consequence of the scaling of contacts is that the basic reproductive number R_0, i.e. the average number of newly infected individuals generated by an infectious one in a fully susceptible population[36], depends on the population size as (see Material and Methods for the complete derivation):

$$R_0(k) = \frac{\lambda}{\mu} N_k^\eta = \frac{\lambda}{\mu} \bar{N}^\eta \frac{k^{(1+\theta)\eta}}{\langle k^{1+\theta} \rangle^\eta} = \mathcal{M} k^\xi. \tag{1}$$

In this expression $\mathcal{M} = \frac{\lambda}{\mu} \frac{\bar{N}^\eta}{\langle k^{1+\theta} \rangle^\eta}$ is a constant that depends on the characteristics of the disease and the metapopulation structure. It is immediate to see from Eq. 1 that $R_0(k)$ will significantly vary from one location to another, depending on the degree of each node and on the exponent $\xi = (1 + \theta)\eta$, which combines the heterogeneity of the traffic flows and of the contact rates.

The necessary and sufficient condition for the local spreading of the disease in nodes of degree k is given by the local epidemic threshold, i.e. $R_0(k) > 1$. It is important to notice that this may not be satisfied in all the subpopulations. Such situation is realistic for a number of epidemic scenarios where, due to specific characteristics of the local population, the value of the basic reproductive number varies across locations[41]. The crucial question in metapopulations systems is evaluating the conditions under which a local epidemic outbreak leads to a global outbreak. This implies defining an invasion threshold R_* for the whole system[17]. In order to find an analytical expression for R_*, we describe the epidemic invasion as a branching process[11,14,15,17,42] relating the number of subpopulations of degree k that have been reached by the epidemic at generation n, D_k^n, with D_k^{n-1}:

$$D_k^n = \sum_{k'} D_{k'}^{n-1} (k' - 1) P(k|k') (1 - R_0(k)^{-\lambda_{k'k}}) \left(1 - \frac{D_k^{n-1}}{V_k}\right). \tag{2}$$

The term $k' - 1$ considers that each diseased subpopulation of degree k' and generation $n - 1$, $D_{k'}^{n-1}$, can seed all the connected nodes but the one from which it received the infection. The term $P(k|k')$ describes the probability that nodes of degree k' are connected with nodes of degree k. We consider uncorrelated networks where this conditional probability does not depend on k' and $P(k|k') = k P(k) / \langle k \rangle$. The term $1 - R_0(k)^{-\lambda_{k'k}}$ defines the probability that, given $\lambda_{k'k}$ infectious individuals seeding a node of degree k, the subpopulation will experience a local outbreak[43]. This number can be estimated as:

$$\lambda_{k'k} = 2 \frac{R_0(k') - 1}{R_0(k')^2} N_{k'} \times \frac{1}{\mu} \times d_{k'k} \times \delta[R_0(k')]. \tag{3}$$

Indeed, the total number of infected individuals generated at the source can be approximated as $2 \frac{R_0(k')-1}{R_0(k')^2} N_{k'}$[17]: infectious individuals recover, on average, after μ^{-1} time steps, and the diffusion rate between the two degree classes is $d_{k'k}$. It is important to notice that such approximations are valid only for $R_0(k') > 1$. Indeed, if this condition is not satisfied the disease will not be able to spread locally in any subpopulation k'. To address this issue, we introduce a step function:

$$\delta[R_0(k')] = \begin{cases} 1 & \text{for } k' | R_0(k') > 1 \\ 0 & \text{for } k' | R_0(k') < 1. \end{cases} \tag{4}$$

Finally, the last term in Eq. 2 represents the fraction of subpopulations of degree k that are not yet infected. By plugging all these terms in Eq. 2, it is possible to solve it analytically and find an explicit expression for the global epidemic threshold:
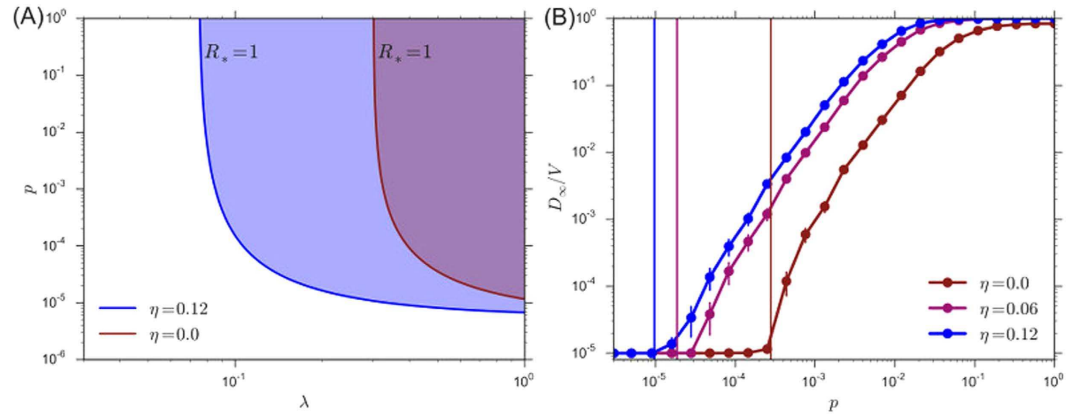
Figure 2. (**A**) Phase diagram defined by the threshold condition $R_*(p, \lambda) = 1$, corresponding to the solid lines, for $\eta = 0$ and $\eta = 0.12$. We consider uncorrelated scale-free networks of $V = 10^5$ nodes, and $P(k) \sim k^{-2.1}$. We set $\theta = 0.5$, $\bar{N} = 10^3$, and $\mu = 0.3$. (**B**) Simulated global attack rate D_∞/V as a function of the mobility rate p for different values of the contact scaling exponent $\eta = 0$, 0.06, 0.12 and $\lambda = 0.35$. Vertical lines indicate the critical threshold value p_c calculated by setting $R_* = 1$ in Eq. 5. Each point is averaged over at least 2×10^3 simulations. Error bars correspond to the standard error of the mean.

$$R_* = \frac{2p\bar{N}}{\mu} \frac{1}{\langle k^{1+\theta} \rangle^2} \left[\langle k^{2+2\theta} \rangle^* - \langle k^{1+2\theta} \rangle^* - \frac{2}{\mathcal{M}} \left(\langle k^{2+2\theta-\xi} \rangle^* - \langle k^{1+2\theta-\xi} \rangle^* \right) \right.$$

$$\left. + \frac{1}{\mathcal{M}^2} \left(\langle k^{2+2\theta-2\xi} \rangle^* - \langle k^{1+2\theta-2\xi} \rangle^* \right) \right] \equiv \frac{2p\bar{N}}{\mu} \mathcal{F}(P(k), \theta, \eta, \lambda, \mu, \bar{N}). \tag{5}$$

See Material and Methods for more details and the Supplementary Information file for the complete derivation. All the moments denoted by a star are calculated over a subset of degree values. More specifically, we define the general starred degree moment as $\langle k^x \rangle^* = \sum_k \delta[R_0(k)] k^x P(k)$. The function $\mathcal{F}$ describes the dependence of the threshold on the properties of the network, the mobility patterns, the scaling of contacts, and the details of the disease. Interestingly, the denominator factor $\langle k^{1+\theta} \rangle^2$ is related to the mobility between subpopulations and not to the spreading dynamics within nodes, therefore the corresponding moment of the degree distribution is calculated over all the values of k.

The expression of the global invasion threshold defines the range of parameters for which a global outbreak is possible, corresponding to the solutions of the equation $R_* = 1$. For $R_* < 1$ an outbreak seeded in any subpopulation will eventually die, while for $R_* > 1$ the contagion process will eventually reach a finite fraction of the system with non-zero probability. The transition between the two regimes is typically continuous[15]. In order to isolate the effects introduced by heterogeneous contact rates, we study the system dynamics for different values of η compared to the case $\eta = 0$ that has been previously studied[15,17]. Indeed, from Eq. 5 it is possible to see that, by setting $\eta = \xi = 0$, we consistently recover the same expression of R_* derived in the case of a constant contact rate across subpopulations[17]. In particular, we compare the value of the critical mobility rate p_c, corresponding to the solution of $R_* = 1$ $\left(p_c = \frac{\mu}{2\bar{N}} \mathcal{F}^{-1} \right)$, in the two cases: $\eta > 0$ and $\eta = 0$. The introduction of a scaling contact rate in every subpopulation, modifies the result of ref. [17] by increasing the overall heterogeneity of the metapopulation system and, eventually, by reducing the critical value of p. More specifically, values of $\eta > 0$ as observed from empirical social networks, alter the spreading dynamics by accelerating the contagion process and thus increasing the value of R_*. This implies that, for a given set of parameters describing the mobility network, the metapopulation system and the transmissibility of the infectious agent, the critical mobility value will be lower for larger values of η. Figure 1A shows the invasion region in the plane $R_*(p, \lambda)$ for $\eta = 0$ and $\eta = 0.12$, with the latter clearly displaying a larger portion of the phase space in the global spreading regime. In particular, the scaling of contacts with subpopulation sizes allows the global spreading of diseases characterized by significantly smaller values of transmissibility λ. As η increases, the difference in the mobility threshold p_c grows smaller with constant λ, while the critical transmissibility λ_c decreases continuously for a given value of p (see Fig. S5 of the Supplementary Information file).

We confirm our analytical findings through extensive numerical simulations performed considering uncorrelated scale-free networks with $V = 10^5$ nodes, and exponent $\gamma = 2.1$[44]. In Fig. 2B, we compare the global attack rate, i.e. the final fraction of subpopulations that experienced a local outbreak, for two identical metapopulation structures and different values of η ($\eta = 0.06$, 0.12). The results of 2×10^3 Monte
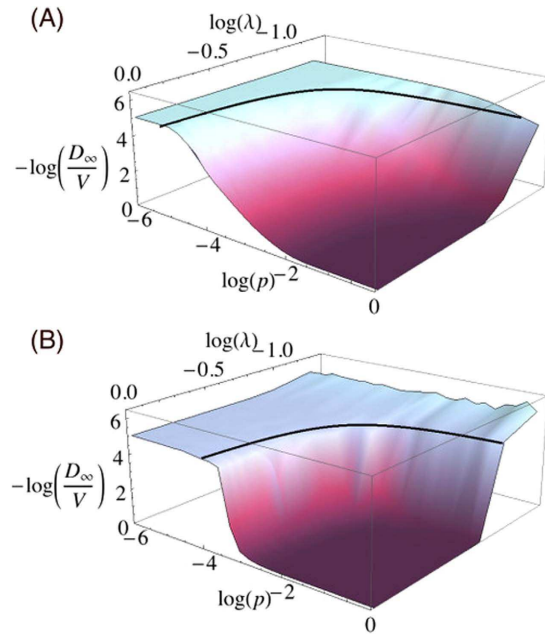
Figure 3. Simulated global attack rate D_∞/V as a two-dimensional function of the mobility rate p and the transmissibility λ for different mobility network structures characterized by $\theta = 0.5$ (A) and $\theta = -0.4$ (B). Black solid lines indicate the analytical predictions for the critical values of p and λ corresponding to $R_* = 1$. Here the network parameters are the same as in Fig. 2 and $\eta = 0.12$. Each point of the phase-space is averaged over 2×10^3 simulations. To facilitate the visual comparison between the simulations and the analytical solutions we plot the z-axis considering the negative $\log_{10}$ of D_∞/V.

Carlo simulations per point show a very good agreement with the theoretical threshold calculated from Eq. 5.

Overall, the global epidemic threshold is determined in a non-linear way, through the exponent ξ, by the interplay between the contact rate heterogeneity, tuned by the exponent η, and the heterogeneity of the mobility patterns, tuned by the exponent θ. The latter can be changed to counterbalance the effect of the contact scaling on the spreading process. In Fig. 3, we show that considering uncorrelated scale-free networks and constant $\eta = 0.12$, higher values of θ correspond to a lower critical mobility rate and a larger invasion regime phase space. On the other side, by assuming a negative value of θ, thus a more homogeneous distribution of the mobility flows across the network, the global spreading regime is suppressed. In both cases, it is remarkable that the numerical simulations show a very good agreement with the theoretical value of the threshold (black solid line in Fig. 3), on the full $(p - \lambda)$ parameter space. Also in this case we considered uncorrelated scale-free networks with $V = 10^5$ nodes, and exponent $\gamma = 2.1$. Each point is averaged in 2×10^3 Monte Carlo simulations. In the Methods Section, we report the full details of the numerical simulations.

It is worth to notice that the value of the global threshold R_*, and, by extension, the value of p_c, depends explicitly on the size of the system, through the moments of the degree distribution of the network. In the regime of large network size and for a range of parameters γ, θ and η that encloses those measured in real systems, it is possible to show that the critical mobility value p_c scales as $p_c(V) \propto V^{\frac{1-\gamma}{2}}$ (see the Supplementary Information file for the full derivation). Interestingly, the threshold vanishes for $V \to \infty$, and its trend explicitly depends only on the exponent γ regulating the heterogeneity of the metapopulation network. Beside the dependence of the global threshold on the network size, our analytical treatment is based on a number of assumptions that may not be satisfied in small size systems. Consequently, by reducing the network size we might expect the numerical invasion threshold to deviate from our theoretical predictions. To test the limits of our treatment, we performed numerical simulations on networks of decreasing size ($V = 10^4$ and $V = 10^3$). Simulation results, shown in Fig. S8 of the Supplementary Information file, indicate that the theoretical predictions are accurate down to the size $V = 10^3$, where finite size effects become more evident.

Discussion

In the present work, prompted by empirical findings, we derived a general framework to study spreading processes in metapopulation systems where the individual contact rates scale with subpopulation sizes.

The scaling properties of social interactions have been derived here and elsewhere[31] from on-line sources and telecommunication datasets. It is not straightforward to assume that such properties would

be observed by the analysis of a large-scale contact network of physical interactions. To date, empirical measures of physical contact networks have been limited to relatively small samples of individuals which do not allow to directly test the scaling hypothesis[26]. There is, however, evidence that a number of properties observed in on-line and telecommunication social networks can be mapped onto the corresponding physical contact network. The correlation between networks inferred by communication platforms and face-to-face interactions has been recently measured in mobile phone data[45,46]. Also, a recent study of contact networks between high-school students found that 67% of the links of their face-to-face contact network, measured with proximity sensors, is present in their Facebook network[47]. Moreover, links of the face-to-face network with aggregate duration larger than a certain threshold correspond all to contacts between Facebook friends[47]. In the case of Twitter, a similar direct empirical comparison is still missing, but recent studies have shown that Twitter mentions reproduce relevant features of real social and mobility networks[30,48–51]. For example, on Twitter individuals devote the large fraction of their communications to a small fraction of ties, i.e. strong ties, and the remaining to occasional contacts, i.e. weak ties[49]. Furthermore, strong ties are statistically localized within the same city[48,51]. Overall, such observations provide indirect evidence supporting the use of Twitter data as a proxy of real social ties relevant for contagion processes. Eventually, it will be important to further confirm the scaling behavior of social contacts by the analysis of additional on-line datasets and, where possible, using real large-scale contact networks.

The effects of local properties of the subpopulations in RD processes, including different local mixing patterns, have been studied in previous works[52–58], but they were generally limited to simplified assumptions on the local contact structure, such as considering only two different contact rates[54,57,58], and by always assuming a constant diffusion rate[53,56,58]. Some recent papers have also considered a power-law distribution of the infectious rates in a metapopulation model[53,59]. However, a comprehensive framework that takes into account the interplay between the heterogeneities of both mobility flows and contact rates was still missing. We have shown that the heterogeneity of the contact rates, introduced by the scaling behavior, promotes the epidemic spreading and such effect is enhanced when the distribution of the mobility flows between subpopulations is heterogeneous, as observed in real mobility networks. Our results represent the first step towards a better analytical understanding of contagion processes in structured subpopulations. The proposed framework can be also extended to include behavioral changes, at the population level, triggered by concerns of infection that might induce a reduction in the contact rates.

Material and Methods

Reciprocal mention interaction (RMI) network of Twitter users. The Twitter dataset was obtained from the raw Twitter Gardenhose feed[60]. The Gardenhose is an unbiased sampling of about 10% of all tweets from Twitter. We considered only Twitter users that were active during an observation period of 8 months, from January 01 2014 to August 31 2014, and built a reciprocal mention interaction (RMI) network between pairs of users, defined as follows: a link is placed between users A and B if and only if user A mentioned user B in one of his/her tweets and user B mention user A back during the observation period.

Geographical mapping of Twitter users. A few percentage of the tweets available through the Gardenhose is provided with GPS information. Based on GPS coordinates, we map a tweet into a geographical area using the following procedure. Typically, an active Twitter user would have a sequence of tweets with GPS information within a given time window. We generate the sequence of locations visited by every Twitter user of our dataset, then we mapped every Twitter user into one geographical area based on his/her most frequently visited location – but only if this represented more than 50% of all locations. For example, let's imagine the Twitter user i tweeted sequentially in city A, A, A, B, A, C, A, A, A and D. In this case, city A has the highest probability of appearance $f_a = \frac{7}{10} = 70\% > 50\%$. We therefore mapped user i into city A and call user i a geo-mappable user. In a different case, a user j tweeted sequentially in city A, A, B, C, D. Although city A has the highest frequency among the locations visited, the relative probability is $f_a = \frac{2}{5} = 40\% < 50\%$ and no city really dominates the geographic distribution of locations visited by user j, thus we do not consider j as a geo-mappable user.

In our work, we mapped the Twitter users into two different types of geographical aggregations: the metropolitan areas of United States and Europe combined (referred to as "metropolitan areas") and the geographical census areas centered around IATA airports[13] (referred to as "basins"). The United States metropolitan areas are defined by the year 2014 United States urban area/urban cluster shapefile from the TIGER/Line database[37]. The European metropolitan areas are defined by the year 2000 morphological zones shapefile (with population larger than 50000). The shapefile can be obtained from the European Union Open Data Portal[38].

The geographical census areas centered around IATA airports are defined[13] by assigning cell of $15' * 15'$ to the closest airport within the same country. The assigning procedure follows a Voronoi-like tessellation[61] with a cut-off scale for the tassels size of $200\,km^2$. After the geo-mapping process, we obtain two subset of users: metropolitan mappable users (MMU) in which each user can be assigned to a metropolitan area in United States or Europe as defined above, or basin mappable users (BMU) in which each user can be assigned to a basin as defined above.

For a given basin/metropolitan area, N is the population, S is the total number of geo-mappable users within the area. The total number of Twitter interactions is $C = \sum_{i \in S_g} c_i$, where c_i is the degree of user i in the subgraph S_g of the entire RMI network. In one case, we compute c_i as the number of interactions of user i confined within the boundary of basin/metropolitan area in the other case, c_i is assumed to be the degree of user i in the entire RMI network. Since the coverage of Twitter user $s = S/N$ differs from one basin/metropolitan area to another, the volume of interaction C_r is rescaled as $C_r = C/s = C\frac{N}{S}$.

Introducing the scaling of contacts into the SIR model.

The SIR is a compartmental model which describes the evolution of a contagious disease in a closed population. The three compartments $S(t)$, $I(t)$, $R(t)$ represent respectively the number of susceptible, infectious and recovered people and the total population $N = S(t) + I(t) + R(t)$ is constant over time.

We assume homogeneous mixing in the population, which means everyone interacts with equal probability with everyone else. Usually it is assumed that the average number of contacts c per individual and unit of time is a constant, $c = c_0$ and does not depend on the population size, then, every susceptible individual has c_0 contacts per unit of time and we define g as the probability of successful disease transmission following a contact. Therefore is convenient to define the transmission rate λ of the disease as[2]:

$$\lambda = -c_0 \log(1 - g) \tag{6}$$

In this framework, it is easy to show that an outbreak of the disease can occur only if $\lambda > \mu$, where μ the recovery rate of the disease. It is a common practice to define the Basic Reproductive Number of the disease as $R_0 = \frac{\lambda}{\mu}$, which has to be greater than 1 for the probability of an outbreak to be larger than 0. In our work, we assumed that the total number of contacts $C = \sum_{i \in S} c_i$ scales super linearly with the population size N:

$$C \propto N^\gamma \quad \text{where} \quad \gamma > 1. \tag{7}$$

So the average per capita contacts rate can be defined as:

$$c \equiv \frac{C}{N} = c_0 N^{\gamma - 1} = c_0 N^\eta \quad \text{where} \quad \eta = \gamma - 1. \tag{8}$$

Consequently the transmission rate will be a function of N:

$$\lambda N^\eta \equiv -c_0 N^\eta \log(1 - g), \tag{9}$$

thus also R_0 will be a function of N:

$$R_0(N) = \frac{\lambda}{\mu} N^\eta. \tag{10}$$

Branching process.

Equation 2 describes the infection dynamics at the level of subpopulations, as a Levins-type process. It can be solved by introducing three main assumptions: the network substrate is uncorrelated, i.e. $P(k|k') = kP(k)/\langle k \rangle$, at the early stage of the epidemic the probability that a subpopulation is not already seeded is almost one, i.e. $\left(1 - \frac{D_k^{n-1}}{V_k}\right) \simeq 1$, and, finally, that $R_0(k) - 1 \ll 1$[17].

In this case, Eq. 2 can be written as:

$$D_k^n = \frac{kP(k)}{\langle k \rangle}(R_0(k) - 1)\sum_{k'}D_{k'}^{n-1}(k' - 1)\lambda_{k'k}. \tag{11}$$

In our framework, it is worth to notice that $R_0(k)$ may be smaller than for some values of k. Therefore, it is reasonable to assume that in a subpopulation with $R_0(k) < 1$ the probability for an outbreak to occur is zero and its contribution to the number of traveling infected individuals, $\lambda_{k'k}$, will be zero. This translates into defining $\lambda_{k'k}$ as:

$$\lambda_{k'k} = \delta[R_0(k')]\alpha(k')N_{k'}d_{k'k}\frac{1}{\mu}, \tag{12}$$

where $\delta[R_0(k')]$ is the Heaviside function of Eq. 4 and the quantity $\alpha(k')N_{k'}$ denotes the total number of infected individuals during the evolution of the epidemic in subpopulation $N_{k'}$. The value of $\alpha(k')$ is dependent on the details of the disease, in particular in the case of $R_0(k') \sim 1$:

$$\alpha\left(k'\right) \simeq \frac{2\left(R_0(k'-1)\right)}{[R_0(k')]^2}.$$

(13)

thus leading to the definition of Eq. 3. By recalling the definition of the diffusion rate, $d_{k'k} = p\frac{\omega_0(k'k)^\theta}{T_{k'}}$, and the stationary value of N_k at equilibrium, $\frac{k^{(1+\theta)}}{\langle k^{1+\theta}\rangle}\overline{N}$, we eventually can write Eq. 11 as:

$$D_k^n = 2\frac{p\overline{N}}{\mu}\left(R_0(k)-1\right)\frac{k^{1+\theta}P(k)}{\langle k^{1+\theta}\rangle^2}\sum_{k'}D_{k'}^{n-1}\frac{(R_0(k')-1)}{[R_0(k')]^2}[(k)^{\prime 1+\theta}-(k')^\theta].$$

(14)

Invasion threshold. To solve the recursive Eq. 14, it is convenient to define the auxiliary function Θ^n as:

$$\Theta^n = \sum_{k'}D_{k'}^n\frac{(R_0(k')-1)}{[R_0(k')]^2}[(k)^{\prime 1+\theta}-(k')^\theta].$$

(15)

In this way Eq. 11 can be conveniently written in the iterative form:

$$\Theta^n = 2\frac{p\overline{N}}{\mu}\frac{1}{\langle k^{1+\theta}\rangle^2}\sum_k\frac{(R_0(k)-1)^2}{[R_0(k)]^2}[k^{2+2\theta}-k^{1+2\theta}]P(k)\Theta^{n-1}.$$

(16)

By expanding the term in $R_0(k)$:

$$\frac{(R_0(k)-1)^2}{[R_0(k)]^2} = 1 - \frac{2}{R_0(k)} + \frac{1}{[R_0(k)]^2}.$$

(17)

and plugging the explicit form $R_0(k) = \mathcal{M}k^\xi$, with $\xi = (1+\theta)\eta$, we have:

$$\Theta^n = \Theta^{n-1}\frac{2p\overline{N}}{\mu}\frac{1}{\langle k^{1+\theta}\rangle^2}\left[\langle k^{2+2\theta}\rangle - \langle k^{1+2\theta}\rangle - \frac{2}{\mathcal{M}}[\langle k^{2+2\theta-\xi}\rangle - \langle k^{1+2\theta-\xi}\rangle]\right.$$
$$\left. + \frac{1}{\mathcal{M}^2}[\langle k^{2+2\theta-2\xi}\rangle - \langle k^{1+2\theta-2\xi}\rangle]\right].$$

(18)

From this equation it is immediate to define the the Global Invasion Threshold R_* as described by Eq. 5 and find the threshold condition on the mobility rate:

$$p\overline{N} \geq \frac{\mu}{2}\langle k^{1+\theta}\rangle^2$$
$$\left[\langle k^{2+2\theta}\rangle - \langle k^{1+2\theta}\rangle - \frac{2}{\mathcal{M}}[\langle k^{2+2\theta-\xi}\rangle - \langle k^{1+2\theta-\xi}\rangle] + \frac{1}{\mathcal{M}^2}[\langle k^{2+2\theta-2\xi}\rangle - \langle k^{1+2\theta-2\xi}\rangle]\right]^{-1}.$$

(19)

Simulation Methods. In all our simulations we consider uncorrelated scale-free networks generated with the UCM algorithm[62]. In particular, we study networks formed by $V = 10^3$, 10^4, 10^5 nodes, and degree distribution $P(k) \sim k^{-2.1}\big|_{k_{min}=2}^{k_{max}=\sqrt{V}}$. Each simulation is started by assigning to each subpopulation the population value at equilibrium N_k, while keeping the average over the whole system constant, $\overline{N} = 10^3$. Furthermore, we seed a randomly selected subpopulation among those with $k = k_{min} = 2$ with 10% of infected individuals. We run the spreading process until the number of infected individuals in the system reaches 0. We consider as diseased any subpopulation in which we observed at least a secondary infection, i.e. an infected seed generates another infected individual. Finally, each simulation point is averaged over 2×10^3 independent simulations. In order to avoid biases associated to specific network structures each simulation is run over a randomly selected network over 60 independent realizations of the UCM algorithm.

References

1. D. Easley & J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press (2010).
2. M. J. Keeling & P. Rohani. *Modeling infectious diseases in humans and animals*. Princeton Univeristy Press (2008).
3. M. E. J. Newman. *Networks, an Introduction*. Oxford University Press (2010).

4. A. Barrat, M. Barthélemy & A. Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press (2008).
5. R. Cohen & S. Havlin. *Complex Networks: Structure, Robustness and Function*. Cambridge University Press, Cambridge (2010).
6. I. Hanski & M. E. Gilplin. *Metapopulation Biology: Ecology, Genetics, and Evolution*. Academic Press, San Diego (1997).
7. R. Pastor-Satorras, C. Castellano, P. Van Mieghem & A. Vespignani. Epidemic processes in complex networks. *Rev. Mod. Phys.* **87,** 925 (2015).
8. L. K. Gallos & P. Argyrakis. Absence of kinetic effects in reaction-diffusion processes in scale-free networks. *Phys. Rev. Lett.* **92,** 138301 (2004).
9. M. D. Holland & A. Hastings. Strong effect of dispersal network structure on ecological dynamics. *Nature.* **456,** 792–795 (2008).
10. H. Nakao & A. S. Mikhailov. Turing patterns in network-organized activator-inhibitor systems. *Nat. Phys.* **6,** 544–550 (2010).
11. P. Bajardi, C. Poletto, J. J. Ramasco, M. Tizzoni, V. Colizza & A. Vespignani. Human mobility networks, travel restrictions and the global spread of 2009 h1n1 pandemic. *PLoS ONE.* **6(1),** e16591 (2011).
12. D. Balcan, V. Colizza, B. Goncalves, H. Hu, J. J. Ramasco & A. Vespignani. Multiscale mobility networks and the large scale spreading of infectious diseases. *Proc. Natl. Acad. Sci. USA.* **106,** 21484 (2009).
13. D. Balcan, H. Hu, B. Goncalves, P. Bajardi, C. Poletto, J. J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Van den Broeck, V. Colizza & A. Vespignani. Seasonal transmission potential and activity peaks of the new inuenza a(h1n1): a monte carlo likelihood analysis based on human mobility. *BMC Med.* **7,** 45 (2009).
14. D. Balcan & A. Vespignani. Phase transitions in contagion processes mediated by recurrent mobility patterns. *Nat. Phys.* **7,** 581–586 (2011).
15. V. Colizza, R. Pastor-Satorras & A. Vespignani. Reaction-diffusion processes and metapopulation models in heterogeneous networks. *Nat. Phys.* **3,** 276 (2007).
16. V. Colizza & A. Vespignani. Invasion threshold in heterogeneous metapopulation networks. *Phys. Rev. Lett.* **99,** 148701 (2007).
17. V. Colizza & A. Vespignani. Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations. *J. Theor. Biol.* **251,** 450–467 (2008).
18. N. M. Ferguson, D. A. T. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn & D. S. Burke. Strategies for containing an emerging inuenza pandemic in southeast asia. *Nature.* **437,** 209 (2005).
19. L. Hufnagel, D. Brockmann & T. Geisel. Forecast and control of epidemics in a globalized world. *Proc. Natl. Acad. Sci. USA.* **101,** 15124 (2004).
20. S. Liu, A. Baronchelli & N. Perra. Contagion dynamics in time-varying metapopulations networks. *Phy. Rev. E,* **87(032805)** (2013).
21. S. Meloni, N. Perra, A. Arenas, S. Gómez, Y. Moreno & A. Vespignani. Modeling human mobility responses to the large-scale spreading of infectious diseases. *Sci. Rep.* **1,** 62 (2011).
22. S. Merler, M. Ajelli, A. Pugliese & N. M. Ferguson. Determinants of the spatiotemporal dynamics of the 2009 h1n1 pandemic in europe: Implications for real-time modelling. *PLoS Comp. Biol.* **7(9),** e1002205 (2011).
23. C. Poletto, M. Tizzoni & V. Colizza. Human mobility and time spent at destination: impact on spatial epidemic spreading. *J. Theor. Biol.* **338,** 41–58 (2013).
24. M. Tizzoni, P. Bajardi, C. Poletto, J. J. Ramasco, D. Balcan, B. Goncalves, N. Perra, V. Colizza, and A. Vespignani. Real-time numerical forecast of global epidemic spreading: case study of 2009 a/h1n1pdm. *BMC Med.* **10,** 165 (2012).
25. H. Hu, K. Nigmatulina & P. Eckhoff. The scaling of contact rates with population density for the infectious disease models. *Math. Biosci.,* **244(2),** 125–134 (2013).
26. C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza & J.-F. Pinton *et al.* Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLoS ONE,* **5(7),** e11596 (2010).
27. J. Mossong, N. Hens, M. Jit, P. Beutels & K. Auranen *et al.* Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.,* **5(3),** e74 (2008).
28. D. Lazer, A. Pentland, L. A. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. A. Christakis, N. Contractor, J. H. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy & M. Van Alstyne. Life in the network: the coming age of computational social science. *Science.* **323(5915),** 721–723 (2009).
29. J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabó & M. Argollo de Menezes *et al.* Analysis of a large-scale weighted network of one-to-one human communication. *New J. Phys.* **9,** 179 (2007).
30. B. Goncalves, N. Perra & A. Vespignani. Modeling users' activity on twitter networks: Validation of dunbar's number. *PLoS ONE.* **6(8),** e22656 (2011).
31. M. Schläpfer, L. M. Bettencourt, S. Grauwin, M. Raschke, R. Claxton, Z. Smoreda, G. B. West & C. Ratti. The scaling of human interactions with city size. *J. R. Soc. Interface.* **11(20130789)** (2014).
32. L. M. Bettencourt, L. Lobo, D. Helbing, C. Kühn & G. B. West. Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl. Acad. Sci. USA.* **104(17),** 7301–7306 (2007).
33. F. Calabrese, D. Dahlem, A. Gerber, D. Paul, C. Xiaoji, J. Rowland, C. Rath & C. Ratti. The connected states of america: Quantifying social radii of inuence. In *Proceedings of IEEE International Conference on Social Computing.* pages 223–230 (2011).
34. L. M. A. Bettencourt. The origins of scaling in cities. *Science* **340,** 1438 (2013).
35. W. Pan, G. Ghoshal, C. Krumme, M. Cebrian & A. Pentland. Urban characteristics attributable to density-driven tie formation. *Nature Com- munications.* **4,** 1961 (2013).
36. R. M. Anderson & R. M. May. *Infectious Diseases in Humans.* Oxford University Press (1992).
37. 2010 census urban lists record layouts. http://www.census.gov/geo/reference/ua/ualists_layout.html.
38. European open data portal. urban morphological zones (2000). http://open-data.europa.eu/data/dataset/xkKFMviSiM3aATMQmSILQ.
39. Socioeconomic data and applications center (sedac), columbia university. gridded population of the world (gpw) v3. http://sedac.ciesin.columbia.edu/gpw.
40. A. Barrat, M. Barthélemy, R. Pastor-Satorras & A. Vespignani. The architeture of complex weighted networks. *Proc. Natl. Acad. Sci. USA.* **101,** 3747–3752 (2004).
41. B. D. Dalziel, K. Huang, J. L. Geoghegan, N. Arinaminpathy, E. J. Dubovi, B. T. Grenfell, S. P. Ellner, E. C. Holmes & C. R. Parrish. Contact heterogeneity, rather than transmission effciency, limits the emergency of canine inuenza virus. *PLoS Pathog.* **10(10),** e1004455 (2014).
42. F. Ball, D. Mollison & G. Scalia-Tomba. Epidemics with two levels of mixing. *Ann. Appl. Probab.* **7,** 46–89 (1997).
43. N. T. Bailey. *The mathematical theory of infectious diseases.* Griffin (1975).
44. M. Catanzaro, M. Boguña & R. Pastor-Satorras. Generation of uncorrelated random scale-free networks. *Phys. Rev. E.* **71(027103)** (2005).
45. F. Calabrese, Z. Smoreda, V. D. Blondel & C. Ratti. Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. *PLoS ONE.* **6(7),** e20814 (2011).
46. J.-P. Onnela, J. Saramäki, J. Hyvonen, G. Szabó & D. Lazer *et al.* Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. USA.* **104(18),** 7332–7336 (2007).
47. R. Mastrandrea, J. Fournet & A. Barrat. Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLoS ONE* **10(9),** e0136497 (2015).

48. P. A. Grabowicz, J. J. Ramasco, B. Gonçalves & V. M. Eguiluz. Entangling mobility and interactions in social media. *PLoS ONE*. **9(3),** e92196 (2014).

49. P. A. Grabowicz, J. J. Ramasco, E. Moro, J. M. Pujol & V. M. Eguiluz. Social features of online networks: The strength of intermediary ties in online social media. *PloS one*, **7(1),** e29358 (2012).

50. Bernardo Huberman, Daniel Romero & Fang Wu. Social networks that matter: Twitter under the microscope. *First Monday*. **14(1)** (Jan 2009).

51. Y. Takhteyev, A. Gruzd & B. Wellman. Geography of twitter networks. *Social networks*. **34(1),** 73–81 (2012).

52. A. Apolloni, C. Poletto & V. Colizza. Age-specific contacts and travel patterns in the spatial spread of 2009 h1n1 inuenza pandemic. *BMC infectious diseases*. **13(1),** 176 (2013).

53. Y.-W. Gong, Y.-R. Song & G.-P. Jiang. Epidemic spreading in metapopulation networks with heterogeneous infection rates. *Physica A*. **416,** 208–218 (2014).

54. H. Lund, L. Lizana & L. Simonsen. Effects of city-size heterogeneity on epidemic spreading in a metapopulation: a reaction-diffusion approach. *J. Stat. Phys*. **151(1),** 367–382 (2013).

55. A. S. Mata, S. C. Ferreira & R. Pastor-Satorras. Effects of local population structure in a reaction-diffusion model of a contact process on metapopulation networks. *Phys. Rev. E*. **88** (2013).

56. C. Shen, H. Chen & H. Zhonghuai. Strategy to suppress epidemic explosion in heterogeneous metapopulation networks. *Phys. Rev. E*. **86,** 036114 (2012).

57. G. Tanaka, C. Urabe & K. Aihara. Random and targeted interventions for epidemic control in metapopulation models. *Sci. Rep*. **4,** 5522 (2014).

58. L. Wang, Z. Wang, Y. Zhang & L. Xiang. How human location-specific contact patterns impact spatial transmission between populations? *Sci. Rep*. **3,** 1468 (2013).

59. T. Takaguchi & R. Lambiotte. Suffcient condition of endemic threshold on metapopulation networks. *J. Theor. Biol*. **380,** 134–143 (2015).

60. Twitter. Twitter public streams (2015). https://dev.twitter.com/streaming/public

61. A. Okabe, B. Boots, K. Sugihara & S. N. Chiu. *Spatial Tessellations - Concepts and Applications of Voronoi Diagrams*. John Wiley (2000).

62. M. Catanzaro, M. Boguña & R. Pastor-Satorras. Generation of uncorrelated random scale-free networks. *Phys. Rev. E*. **71,** 027103 (2005).

Acknowledgements

Author Contributions

M.T. and N.P. conceived the research, M.T., N.P. and D.B. performed simulations and analyzed the results. K.S. and M.K. collected and analyzed the data. M.T. and N.P. wrote the manuscript. All authors reviewed and approved the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Tizzoni, M. *et al.* The Scaling of Human Contacts and Epidemic Processes in Metapopulation Networks. *Sci. Rep.* **5**, 15111; doi: 10.1038/srep15111 (2015).